РАЗРАБОТКА МЕТОДОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ СЕТЕВЫХ РЕСУРСОВ ИНФОРМАЦИОННЫХ СИСТЕМ

Вишняков А.С.[1], Макаров А.Е.[2], Уткин А.В.[3], Зажогин С.Д.[4], Бобров А.В.[5]
Email: Vishniakov1158@scientifictext.ru

[1]Вишняков Александр Сергеевич – ведущий инженер,
системный интегратор «Крастком»;
[2]Макаров Анатолий Евгеньевич – архитектор решений,
Российская телекоммуникационная компания «Ростелеком»,
г. Москва;
[3]Уткин Александр Владимирович – старший инженер,
Международный системный интегратор «EPAM Systems», г. Минск, Республика Беларусь;
[4]Зажогин Станислав Дмитриевич - старший разработчик,
Международный IT интегратор «Hospitality & Retail Systems»;
[5]Бобров Андрей Владимирович – руководитель группы,
группа технической поддержки,
Компания SharxDC LLC,
г. Москва

Аннотация: *рассмотрены методы нечеткой кластеризации и сделаны выводы по их применению в информационных системах. Показано, что метод нечетких с-средних обладает устойчивостью с точки зрения использования естественного нечеткого классификатора. Был рассмотрен метод нечеткой кластеризации k-средних и метод энтропии, показано, что при кластеризации информационных систем метод нечетких с-средних обладает большей устойчивостью. Рассмотрены возможности применения в методах нечетких с-средних метрики Махаланобиса, в частности были представлены алгоритмы Густафсона-Кесселя и Кульбака-Лейблера. Показана необходимость при построении нелинейных границ кластеров использования ядерных методов кластеризации. Разработан комплексный алгоритм определения оптимального способа для кластеризации элементов информационной системы.*

Ключевые слова: *информационные системы, метод нечетких с-средних, метод энтропии, алгоритм Густафсона-Кесселя, алгоритм Кульбака-Лейблера, ядерные методы кластеризации.*

DEVELOPMENT OF METHODS FOR FUZZY CLUSTERING OF NETWORK INFORMATION SYSTEMS

Vishniakov A.S.[1], Makarov A.E.[2], Utkin A.V.[3], Zazhogin S.D.[4], Bobrov A.V.[5]

[1]Vishniakov Alexandr Sergeevich – Lead System Engineer,
SYSTEM INTEGRATOR «KRASTCOM»;
[2]Makarov Anatoly Evgenevich – Solutions Architect,
ROSTELECOM INFORMATION TECHNOLOGY,
MOSCOW;
[3]Utkin Alexander Vladimirovich – Senior Engineer,
INTERNATIONAL SYSTEM INTEGRATOR EPAM SYSTEMS, MINSK, REPUBLIC OF BELARUS;
[4]Zazhogin Stanislav Dmitrievich – Senior Software Engineer,
International IT Integrator Hospitality & Retail Systems;
[5]Bobrov Andrei Vladimirovich – Team leader,
TECHNICAL SUPPORT GROUP,
SHARXDC LLC,
MOSCOW

Abstract: *the methods of fuzzy clustering are considered and conclusions on their use in information systems are made. It is shown that the method of fuzzy c-means is stable in terms of using a natural fuzzy classifier. The fuzzy clustering method of k-means and the entropy method were considered, and it was shown that when information systems are clustering, the fuzzy c-means method is more stable. The possibilities of using fuzzy methods with the Mahalanobis metrics are considered, in particular, the Gustafson-Kessel and Kullback–Leibler algorithms were presented. The necessity of using kernel clustering methods in the construction of nonlinear cluster boundaries is shown. A complex algorithm for determining the optimal method for clustering information system elements has been developed.*

Keywords: *information systems, fuzzy c-means method, entropy method, Gustafson-Kessel algorithm, Kullback–Leibler algorithm, kernel clustering methods.*

УДК 331.225.3

Введение

Кластерный анализ данных путем автоматической генерации групп объектов информационных систем на основании параметров, определяющих их сходство, широко используется в области современных информационных технологий [1, 4-10]. Среди большого количества перспективных методов кластеризации необходимо выделить группу алгоритмов, которые основываются на методе нечеткой кластеризации *c*-средних, преимуществом которых является простота реализации и устойчивость, что определяет ***актуальность*** данного исследования.

Анализ последних исследований и публикаций в данной области показал перспективность метода нечеткой кластеризации *k*-средних [4, 5] и метода энтропии [6, 7], которые могут быть взяты за основу при разработке более сложных алгоритмов. В частности при развитии метода нечеткой кластеризации *c*-средних на метрику Махаланобиса могут быть использованы алгоритмы Густафсона-Кесселя [7] и Кульбака-Лейблера [8]; а для построения нелинейных границ кластеров актуально использовать ядерные методов кластеризации [9, 10]. Проведенный анализ также показал на отсутствие целостной методологии нечеткой кластеризации информационных систем, что было выделено как нерешенную часть общей проблемы.

Целью работы, таким образом, стала разработка комплексного алгоритма определения оптимального способа для нечеткой кластеризации элементов информационной системы, что позволяет определить эффективность метода через решение математической задачи определения экстремума целевой функции.

1. Применение метода нечеткой кластеризации *c*-средних в информационных системах

Метод кластеризации *c*-средних (*c*-means) в общем виде рассматривает процесс кластеризации как выделение центров кластеров $c_k \in [c_1; c_K]$ для набора объектов $x_j \in [x_1; x_J]$ через функцию u_k^j, которая определяется через соответствующую метрику (рис. 1). В случае применения метода кластеризации *c*-средних обычно используется метрика Минковского для $n = 2$ (т.е. Евклидова метрика).

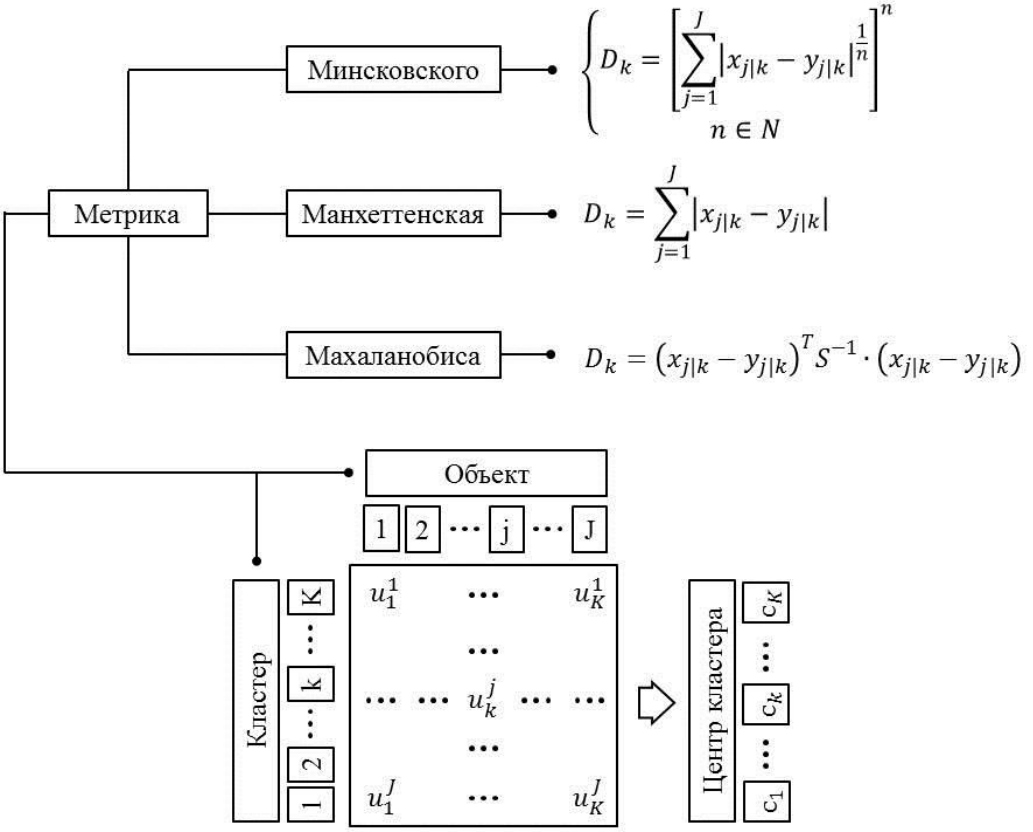

Рис. 1. Обобщенная схема метода кластеризации c-средних

При этом всё множество значений, которые принимает функция u_k^j описывается через следующую систему уравнений [2, 3]:

$$\begin{cases} \begin{cases} u_k^j \in [0;1] \\ j \in [1;J] \\ k \in [1;K] \end{cases} \\ \begin{cases} \sum_{j=1}^{J} u_k^j = 1 \\ k \in [1;K] \end{cases} \end{cases} . \qquad (1)$$

Наиболее простым в описании и реализации является метод четкой кластеризации *c*-средних (CCM: crisp *c*-means), алгоритм которого представлен на рис. 2. Данный метод включает в себя создание случайных центров кластеризации, соотнесении множества объектов с данными центрами кластеризации в соответствии с выбранной метрикой, расчет новых центров кластеризации как центроидов (центров масс) множества. В случае сходимости центроидов алгоритм CCM считается завершенным.

Рис. 2. Схема реализации алгоритма четкой кластеризации c-средних

Центры кластеризации в алгоритме CCM определяются как [2, 3]:

$$\begin{cases} c_k = \frac{\sum_{j=1}^{J} x_j}{n_k} \\ x_j \in \{c_k\} \end{cases} , \qquad (2)$$

где n_k — количество объектов в кластере k.

Метод нечеткой кластеризации *c*-средних (FCM: fuzzy *c*-means) аналогичным образом может быть определен через уравнение для целевой функции G:

$$\begin{cases} G = \sum_{k=1}^{K} \sum_{j=1}^{J} \left(u_k^j\right)^m D\left(x_j, c_k\right) \\ m \in [1, \infty) \end{cases} , \qquad (3)$$

где $D(x_j, c_k)$ — расстояние, которое определяется соответствующей метрикой. Данный метод включает в себя минимизацию функции G относительно множеств $\{u_k^j\}$ и $\{c_k\}$ с последующей проверкой сходимости полученных значений (рис. 3)

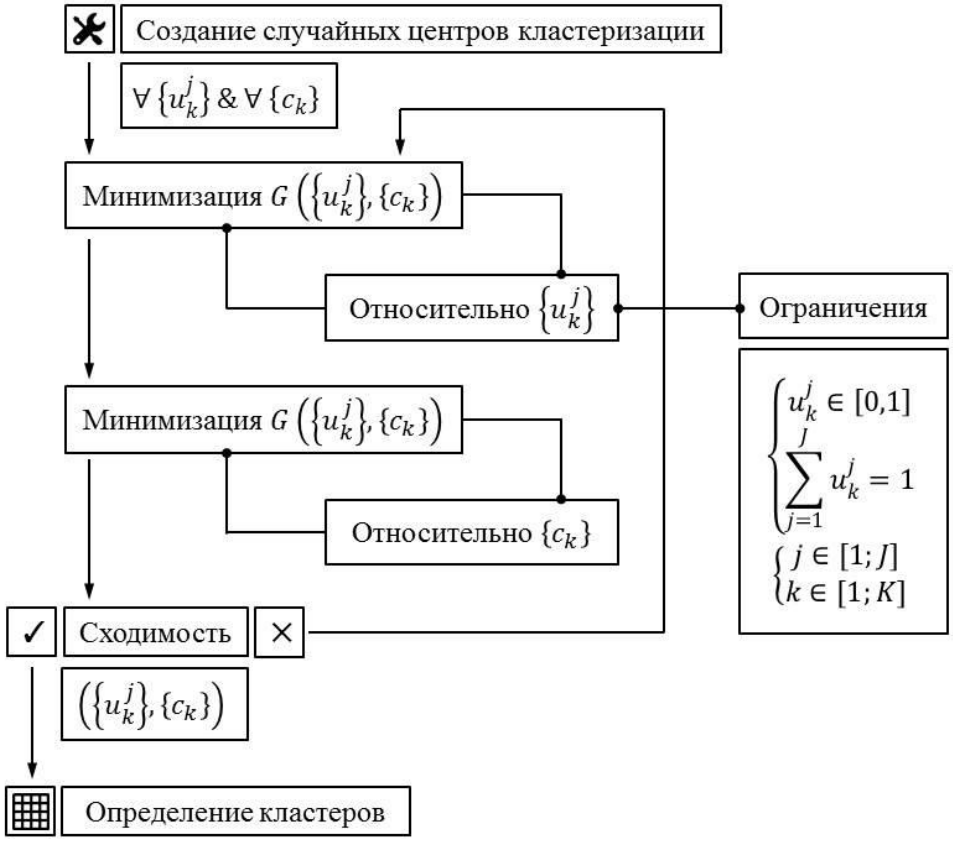

Рис. 3. Схема реализации алгоритма нечеткой кластеризации с-средних

Соответственно, при минимизации функции G относительно множества $\{u_k^j\}$ вносятся следующие ограничения.

$$\begin{cases} \begin{cases} u_k^j \in [0,1] \\ \sum_{j=1}^{J} u_k^j = 1 \end{cases} \\ \begin{cases} j \in [1;J] \\ k \in [1;K] \end{cases} \end{cases}. \qquad (4)$$

Как можно увидеть, при $m = 1$ метод FCM переходит в форму CCM, в то время как для $m > 1$ оптимальные значения целевых функций u_k^j и c_k ($\overline{u_k^j}$ и $\overline{c_k}$) могут быть определены через систему уравнений для всех x_j, что не являются центрами кластеризации:

$$\begin{cases} \overline{u_k^j} = 1/\sum_{k=1}^{K} \sqrt[m-1]{\frac{D(x_j,\bar{c}_k)}{D(x_j,\bar{c}_k)}} \\ \bar{c}_k = \frac{\sum_{j=1}^{J}\left(\left(u_k^j\right)^m x_j\right)}{\sum_{j=1}^{J}\left(u_k^j\right)^m} \\ x_j \neq c_k \end{cases} . \quad (5)$$

Соответственно, для x_j, которые являются центрами кластеризации может быть определена следующая система уравнений:

$$\begin{cases} \overline{u_k^j} = 1/\sum_{k=1}^{K} \sqrt[m-1]{\frac{D(x_j,c_k)}{D(x_j,c_{k\prime})}} \\ \bar{c}_k = \frac{\sum_{j=1}^{J}\left(\left(u_k^j\right)^m x_j\right)}{\sum_{j=1}^{J}\left(u_k^j\right)^m} \\ x_j = c_k \\ c_{k\prime} \neq c_k \end{cases} . \quad (6)$$

Системы уравнений (5) и (6) можно рассматривать как основу построения алгоритмов кластеризации элементов информационных систем на основе ССМ.

2. Особенности развития методологии нечеткой кластеризации информационных систем

Наиболее популярным методом кластеризации, который в какой-то мере обобщает приведенный выше математический аппарат, является метод *k*-средних, который определяется через ближайший центр распределения при помощи четкого классификатора U_i [4, 5]:

$$\begin{cases} U_i^K(x;c) = 1 \\ i = arg\left(min_k\left(D(x;c_k)\right)\right) \\ k \in [1;K] \end{cases} . \quad (7)$$

Соответственно для двух областей V_i и V_i' может быть определено $U_k(x;V) = 0$ и $U_i(x;V) > U_k(x;V)$ соответственно для любого $j \neq k$

Также можно предположить применение целевой функции нечетких средних на основе метода энтропии [6, 7]:

$$\begin{cases} G^{ent} = \sum_{k=1}^{K}\sum_{j=1}^{J}\left\{u_k^j \cdot D\left(x_j,c_k\right) + \lambda^{-1}u_k^j \cdot \log\left(u_k^j\right)\right\} \\ \lambda > 0 \end{cases} . \quad (8)$$

Аналогично схеме представленной на рис. 3 в результате минимизации функции можно получить значения

$$\begin{cases} u_k^j = \frac{e^{-\lambda D\left(x_j, c_k\right)}}{\sum_{j=1}^{J} e^{-\lambda D\left(x_j, c_k\right)}} \\ c_k = \frac{\sum_{j=1}^{J} u_k^j x_J}{\sum_{j=1}^{J} u_k^j} \end{cases}. \qquad (9)$$

Тогда для данного метода классификатор может быть определен как

$$U_i^{ent} = \frac{e^{-\lambda D(x, c_k)}}{\sum_{j=1}^{J} e^{-\lambda D(x, c_k)}}. \qquad (10)$$

Сравнение FCM, метода *k*-средних и метода энтропии возможно через соотнесение функций $U_i(x;c)$, $U_i^K(x;c)$ и U_i^{ent}. Если x достаточно далеко от центров $c_k \in [c_1; c_K]$ для FCM $U_i(x;c) \approx 1/K$, соответственно $U_i^K(x;c) = 1$ и $U_k^K(x;c) = 0$ для $k \neq i$. Для метода энтропии, в свою очередь, $U_i^{ent} \approx 1$. Следовательно метод *k*-средних и метод энтропии в значительной степени зависят от объектов, которые находятся далеко от центров кластеризации. Более того, функция $U_i(x;c)$ характеризуется максимумом в точке $x = c_k$. Стандартно FCM при кластеризации объектов информационных систем имеет преимущества перед методом *k*-средних и методом энтропии.

3. Обобщение метода нечеткой кластеризации *c*-средних для метрики Махаланобиса

Значительное число методов, которые базируются на методе нечеткой кластеризации *c*-средних, основываются на метрике Махаланобиса, расстояние для которой определяется как меру несходства между двумя векторами [8]:

$$D_k = (x_k - y_k)^T S^{-1} \cdot (x_k - y_k), \qquad (11)$$

где S — матрица ковариации.

Для алгоритма Густафсона-Кесселя включает кластерные ковариационные переменные $S_k \in [S_1; S_K]$. Т.о., целевая функция может быть определена как:

$$\begin{cases} G\left(\{u_k^j\}, \{c_k\}, S\right) = \sum_{k=1}^{K} \sum_{j=1}^{J} \left(u_k^j\right)^m D\left(x_j, c_k; S_k\right) \\ m > 1 \end{cases}. \qquad (12)$$

Также целевую функцию можно выразить через дополнительный параметр α_k, набор которых формирует матрицу $A = \{\alpha_1, \ldots, \alpha_K\}$:

$$\begin{cases} G\left(\{u_k^j\}, \{c_k\}, S, A\right) = \sum_{k=1}^{K} \sum_{j=1}^{J} (\alpha_k)^{1-m} \left(u_k^j\right)^m D\left(x_j, c_k; S_k\right) \\ m > 1 \end{cases}, \qquad (13)$$

где α_k определяется через следующую систему уравнений:

$$\begin{cases} \sum_{k=1}^{K} \alpha_k = 1 \\ \alpha_k \geq 0 \\ k \in [1;K] \end{cases} . \qquad (14)$$

Т.о., в данном случае необходимо провести минимизацию четырех функций u_k^j, , c_k, S_k и α_k. Аналитическое решение в общем случае может быть представлено в следующем виде:

$$\begin{cases} u_k^j = 1 / \sum_{j=1}^{J} \left(\sqrt[m-1]{\frac{D(x_j, c_k; S_k)}{D(x_j, c_k; S_k)}} \right) \\ c_k = \frac{\sum_{j=1}^{J} \left(u_k^j\right)^m x_j}{\sum_{j=1}^{J} \left(u_k^j\right)^m} \\ S_k \sim \frac{1}{\hat{S}_k} \sum_{j=1}^{J} \left(u_k^j\right)^m \left(x_j - c_k\right)\left(x_j - c_k\right)^T \\ \alpha_k = 1 \Big/ \sum_{j=1}^{J} \left(\sqrt[m]{\frac{\sum_{j=1}^{J} \left(u_k^j\right)^m D(x_j, c_k; S_k)}{\sum_{j=1}^{J} \left(u_k^j\right)^m D(x_j, c_k; S_k)}} \right) \end{cases}, \qquad (15)$$

где $\hat{S}_k$ определяется как

$$\hat{S}_k = \sum_{j=1}^{J} \left(u_k^j\right)^m \left(x_j - c_k\right)\left(x_j - c_k\right)^T. \qquad (16)$$

Соответственно сходимость алгоритма Густафсона-Кесселя вычисляется через оптимизацию четырех функций.

Другим вариантом метода нечеткой кластеризации *c*-средних для метрики Махаланобиса актуальным для информационных систем является метод Кульбака-Лейблера. Следует отметить, что данный подход включает в себя энтропийный алгоритм рассмотренный выше. Целевая функция для метода Кульбака-Лейблера может быть определена как

$$G_{KL} = \sum_{k=1}^{K} \sum_{j=1}^{J} u_k^j D\left(x_j, c_k; S_k\right) + \sum_{k=1}^{K} \sum_{j=1}^{J} \left(v \cdot u_k^j \cdot log\left(\frac{u_k^j}{\alpha_k}\right) + log|S_k| \right).$$

(17)

Соответственно для метода Кульбака-Лейблера как и в случае метода Густафсона-Кесселя производится минимизация четырех функций u_k^j, , c_k, S_k и α_k с аналитическим решением, которое может быть представлено в виде:

$$\begin{cases} u_k^j = \frac{\alpha_i}{|S_i|} \cdot e^{-\frac{D(x_j,c_k,S_k)}{v}} \Big/ \sum_{j=1}^{J} \left(\frac{\alpha_i}{|S_i|} \cdot e^{-\frac{D(x_j,c_k,S_k)}{v}} \right) \\ c_k = \frac{\sum_{k=1}^{N} u_k^j \cdot x_j}{\sum_{k=1}^{N} u_k^j} \\ S_k = \frac{1}{\sum_{j=1}^{J} u_k^j} \sum_{j=1}^{J} u_k^j \cdot (x_j - c_k)(x_j - c_k)^T \\ \alpha_k = \frac{1}{K} \sum_{j=1}^{J} u_k^j \end{cases}, \qquad (18)$$

По уравнениям (17-18) можно увидеть, что метод Кульбака-Лейблера соответствует статистической модели, известно как смешанная модель Гаусса. Целевые функции по методу Кульбака-Лейблера рассчитываются проще, чем по методу Густафсона-Кесселя, но при этом метод Густафсона-Кесселя обладает большей устойчивостью.

4. Применение в информационных системах ядерных методов нечеткой кластеризации *c*-средних

Ядерные алгоритмы нечеткой кластеризации *c*-средних большей частью основываются на методах опорных векторов и ядерных функциях. Причина, по которой имеет смысл использовать ядерные методы для кластеризации, заключается в том, что *k*-средние и нечеткие *c*-средние характеризуются линейными границами между кластерами областей, в то время как для более гибкого подхода необходимо использовать нелинейные границы [9, 10].

В рамках данного подхода набор объектов $x_j \in [x_1; x_J]$ может быть представлен в виде многомерного отображения $\Phi_j \in [\Phi_1; \Phi_J]$, при этом ядерная функция представляется в гильбертовом пространстве:

$$K(x, y) = \langle \Phi(x_i), \Phi(x_j) \rangle_H . \qquad (19)$$

В ядерных алгоритмах нечеткой кластеризации *c*-средних, таким образом, целевая функция использует набор $\Phi(x_j) \in [\Phi(x_1); \Phi(x_J)]$ и центрами кластеризации в гильбертовом пространстве $c_k^H \in [c_1^H; c_K^H]$:

$$\begin{cases} G(\{u_k^j\}, \{c_k\}) = \sum_{k=1}^{K} \sum_{j=1}^{J} (u_k^j)^m \left\| \Phi(x_j) - c_k^H \right\|_H^2 . \\ m > 1 \end{cases} \qquad (20)$$

Соответственно при определении сходимости целевой функции необходимо минимизировать $\{u_k^j\}$ и функцию, определяющую центры кластеризации в гильбертовом пространстве:

$$\begin{cases} u_k^j = 1 \Big/ \sum_{j=1}^{J} \left(\sqrt[m-1]{\frac{\left\| \Phi(x_j) - c_k^H \right\|_H^2}{\left\| \Phi(x_j) - c_k^H \right\|_H^2}} \right) \\ c_k^H = \sum_{j=1}^{J} \left((u_k^j)^m \Phi(x_j) \right) \Big/ \sum_{j=1}^{J} (u_k^j)^m \end{cases} . \qquad (21)$$

Т.о., для определения центров кластеризации в рамках данного подхода нужно либо определить функцию $\Phi(x_j)$, либо исключить ее из описания ядерной функции.

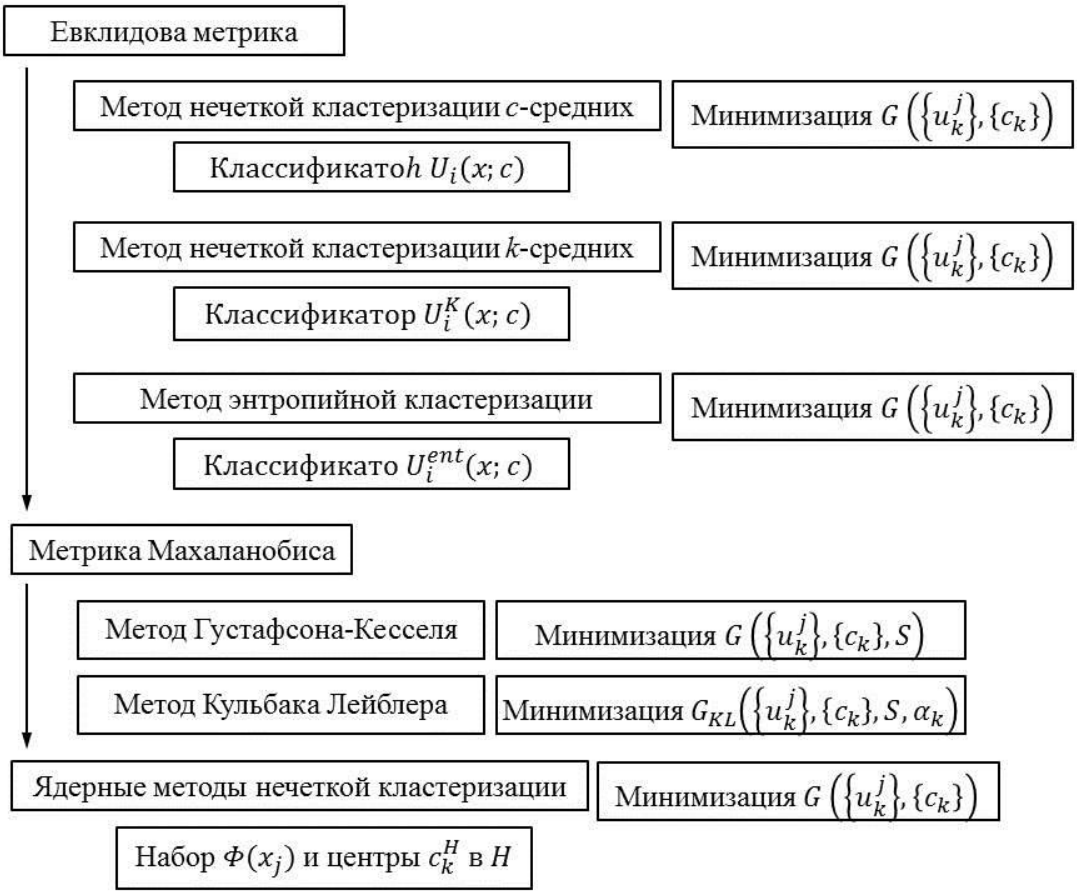

Рис. 4. Алгоритм определения оптимального способа нечеткой кластеризации элементов информационной системы

Проведенный анализ позволяет построить комплексный алгоритм автоматического определения оптимального способа кластеризации элементов информационной системы, который базируется на методах нечеткой кластеризации *c*-средних (рис. 4).

Выводы

В результате проведенного анализа были изучены методы нечеткой кластеризации и сделаны выводы по их применения в информационных системах. Метод нечетких *c*-средних обладает устойчивостью с точки зрения использованием естественного нечеткого классификатора. В рамках данного анализа были рассмотрены метод нечеткой кластеризации *k*-средних и метод энтропии, было показано, что метод *k*-средних и метод энтропии в значительной степени зависят от объектов, которые находятся далеко от центров кластеризации, т.о. для кластеризации информационных систем в большей степени подходит метод нечетких *c*-средних. В качестве развития данного подхода на метрику Махаланобиса были представлены алгоритмы Густафсона-Кесселя и Кульбака-Лейблера. Также было показана необходимость использования ядерных методов кластеризации, которые характеризуются нелинейными границами кластеров областей для применения более гибкого подхода в этой области. В результате был разработан комплексный алгоритм определения оптимального способа для кластеризации элементов информационной системы.

Список литературы / References

1. *Miyamoto S., Ichihashi H., Honda K.* Algorithms for Fuzzy Clustering, Springer. Berlin, 2008.
2. *Bezdek J.C.* Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.
3. *Girolami M.* Mercer kernel based clustering in feature space, IEEE Trans. on Neural Networks. Vol. 13. № 3. Pp. 780–784, 2002.
4. *Haqiqi B.N. & Kurniawan R.,* 2015. Analisis Perbandingan Metode Fuzzy C-Means Dan Subtractive Fuzzy C-Means. Media Statistika, 8(2). doi:10.14710/medstat.8.2.59-67.

5. *Lee S., Kim J. & Jeong Y.*, 2017. Various Validity Indices for Fuzzy K-means Clustering. Korean Management Review, 46(4), 1201-1226. doi:10.17287/kmr.2017.46.4.1201.
6. *Kanzawa, Y., Endo Y. & Miyamoto S.,* 2008. Fuzzy classification function of entropy regularized fuzzy c-means algorithm for data with tolerance using kernel function. 2008 IEEE International Conference on Granular Computing. doi:10.1109/grc.2008.4664765.
7. *Yasuda M.,* 2014. Q-increment deterministic annealing fuzzy c-means clsutering using Tsallis entropy. 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). doi:10.1109/fskd.2014.6980802.
8. *Chen S.,* 2017. An improved fuzzy decision analysis framework with fuzzy Mahalanobis distances for individual investment effect appraisal. Management Decision, 55(5), 935-956. doi:10.1108/md-11-2015-0512.
9. *Cai,Q.,& Liu,W.,* 2009. TSK fuzzy model using kernel-based fuzzy c-means clustering. 2009. IEEE International Conference on Fuzzy Systems. doi:10.1109/fuzzy.2009.5277146
10. *Baili N.,* 2013. Unsupervised and semi-supervised fuzzy clustering with multiple kernels. Louisville, KY: University of Louisville.